Variables.

Variable is a property that can take any values.

Types of variable:

- Quantitative variable : Numerically measured variable. Ex- Age, weight, Distance.
- Qualitative variable: Categarical variable, they are grouped together based on same characteristics. Ex - Gender, Income gap.

Quantitative variable.

Discrete variable

[Non - negative no.s only whale no.s]

Ex- No.s of Book, No.s of Acc/no one can hald.

Continuous Variable

[Continuous no.s even decimal or negative no.s]

Ex- Age, weight Speed.

Measure of central tendency (C.T):

- It's a single value that attempts to describe a set of data identifying the central position.

- Mean
- Median
- Mode.

Mean : It's average of the data.

Population = N (no.s of Population)

Population mean $= \mu = \sum_{i=1}^{N} \frac{x_i}{N}$

Sample = n (no.s of sample)

Sample mean $= \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

$N >> n$

Median : It's central number after sorting the data.

- If no.s of elements are even we find average of central elements.

Ex - median of [2, 6, 10, 12, 16, 17]

$$\text{median} = \frac{10+12}{2} = 11$$

- If no.s of element are **odd** we **select** the **central number**.

Ex - median of [1, 2, 6, 10, 12, 16, 17]

median = 10

Mode : **Most frequent** occuring element.

Ex - Mode of [4, 6, 4, 8, 10, 6, 9, 4, 5]

Mode = 4

Note: when we have outlier in data we use median insted of Mean.

Ex - Age - $[2, 4, 6, 12, 18, 20, \underbrace{86, 97}_{\text{outlier.}}]$

Median = 15

Mean = 30

with outlier data - Median

No outlier data - Mean.

Measure of dispersion:

- **Variance** (σ^2)

It refers to statistical measurement of the spread b/w numbers in a data set. Specifically it measures how far each no.s is from mean.

Variance

Population variance (σ^2)

Sample variance (s^2)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

N - No. of population

n - No. of sample

μ - Population mean

$\bar{x}$ - Sample mean.

Ex-

data1 = [2, 4, 3, 7]

$\mu = 4$

$$\sigma^2 = \frac{(2-4)^2 + (4-4)^2 + (3-4)^2 + (7-4)^2}{4}$$

$$= \frac{4 + 0 + 1 + 9}{4}$$

$$= 3.5$$

$\text{data 2} = [2, 4, 6, 10, 11, 13, 18, 19, 21, 26]$

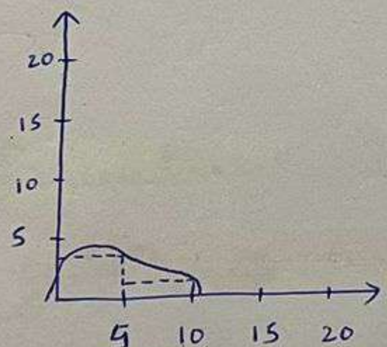
$\mu = 13$

$$\sigma^2 = \frac{(13-2)^2 + (13-4)^2 + (13-6)^2 + (13-10)^2 + (13-11)^2 + (13-13)^2 + (13-18)^2 + (13-19)^2 + (13-21)^2 + (13-26)^2}{10}$$

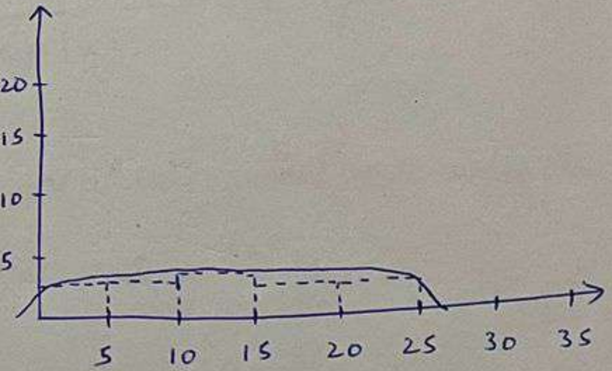
$$= \frac{121 + 81 + 49 + 9 + 4 + 0 + 25 + 36 + 64 + 169}{10}$$

$$= 55.8$$

Plotting bath on histogram.



data 1



data 2

$$\sigma^2_{data2} > \sigma^2_{data1}$$

$$(\text{Distribution})_{data2} > (\text{distribution})_{data1}$$

⇒ with increase in variance spread keep increasing the distribution.

- **Standard deviation (σ):**

It's sq. root of variance (σ^2). It's measure of the amount of variation or dispersion of a set of values.

St. deviation of (1, 2, 3, 4, 5) is

$$\sigma^2 = 2$$
$$\sigma = 1.414$$

Percentile and Quartiles.

$$\text{Percentage} = \frac{\text{Occurance}}{\text{Tatal}} \times 100$$

Percentile: It's a value below which Certain percentage of value lies.

Ex- 75 percentile means the performance is better than 75% of entire population.

Quartile:

It's a type of quantile which divides the number of data points into four parts or quaters.

5 number Summary.

- Minimum
- First quartile (Q_1) 25 percentile
- Median
- Third quartile (Q_3) 75 percentile.
- Maximum.

Ex- Data - [2, 4, 6, 7, 9, 11, 13, 19]

$$Q_1 = \frac{25}{100} \times n = \frac{25}{100} \times 8$$ [n = no. of datapoint]

$$= 2^{nd} = 4$$

$$Q_3 = \frac{75}{100} \times n = \frac{75}{100} \times 8$$

$$= 6^{th} = 11$$

Inter Quartile Range (IQR) = $Q_3 - Q_1$

$$= 11 - 4 = 7$$

lower fench $= Q_1 - 1.5 \times (IQR)$

$$= 4 - 1.5 \times 7$$

$$= -6.5$$

Higher fench $= Q_3 + 1.5 \times (IQR)$

$$= 11 + 1.5 \times 7$$

$$= 21.5$$